A NEW ESTIMATOR FOR THE NUMBER OF SPECIES IN A POPULATION

BY L. CECCONI, * A. GANDOLFI * AND C.C.A. SASTRI

University of Firenze and Missouri University of Science and Technology

We consider the classic problem of estimating T, the total number of species in a population, from repeated counts in a simple random sample and look first at the Chao-Lee estimator: we initially show that such estimator can be obtained by reconciling two estimators of the unobserved probability, and then develop a sequence of improvements culminating in a Dirichlet prior Bayesian reinterpretation of the estimation problem. By means of this, we obtain simultaneous estimates of T, the normalized interspecies variance γ^2 and the parameter λ of the prior. Several simulations show that our estimation method is more flexible than several known methods we used as comparison; the only limitation, apparently shared by all other methods, seems to be that it cannot deal with the rare cases in which $\gamma^2 > 1$.

1. Introduction. We consider the classic problem of estimating the number T of species in a population, and, subsequently, their distribution, from a simple random sample drawn with replacement. We are interested in the "small sample" regime in which it is likely that not all species have been observed. Problems of this kind arise in a variety of settings: for example, when sampling fish from a lake or insects in a forest (see, for instance, Shen, Chao and Lin (2003) [47] on how to use estimates of T to predict further sampling, or [7]); or when estimating the size of a particular population (see [6]); or when trying to guess how many letters an alphabet or how many specific groups of words a language contains (see [14]) or how many words a writer knows (see [19]); or, even, when determining how many different coins were minted by an ancient population (Esty [21]). Because of its great interest this has become a classic in probability, and there has been a great number of studies suggesting methods for the estimation of T. See, for instance, [8] for a review through 1993, [23] for some further details and Colwell's Estimates for software implementing a large number of estimators.

1

In particular, [8] calls for some development of the Bayesian method for the estimation of T, which is the direction that we eventually have taken.

In this paper we start, in fact, by analyzing one well known estimator of T, namely the one by Chao and Lee ([13]). One of our results shows that the estimator can be obtained by reconciling two estimators of the unobserved probability U: one being an extended version of Laplace's "add λ" ([34]) and the other the estimator by Turing and Good ([24]), provided that the normalized interspecies variance γ^2 is interpreted as the inverse of the λ. Then we proceed by developing simultaneous methods for estimating T and λ (or γ^2, which is the same).

By such methods we improve on the original Chao-Lee estimation, but the estimators we obtain are shown by simulations to have some serious defects. It is for this reason that we perform a more fundamental analysis of the problem by means of a Bayesian approach. This is based on a Dirichlet prior with parameter λ on the probabilities of T species (see [33], [32], [25], and [49] for an historical description); the parameter turns out to be the same as the one in Laplace's method. The simultaneous estimation that we develop now takes into account a posterior second moment of the random species probabilities compared to the classical Good Toulmin estimator for the same quantity (see [27]).

Let us mention that the empirical Bayesian approach used here is different from that of existing results in the literature. The method in [41] is, in fact, limited to uniform species distributions. On the other hand, the general Bayesian approach in Boender and Rinnoy Kan (1987) [4] starts from a prior distribution of T and, conditionally to T, a uniform or Dirichlet(λ) prior on the species probability, but then introduces a (level III) prior on λ itself (as suggested in [26]) which in turn requires the introduction of a further parameter (Boender and Rinnoy Kan (1987) [4], formulae (10) and (11)), with then no analytical expression for the posteriors. In the end, this direction seems to include several undetermined choices (the prior on T and the extra parameter at level III) and no simple analytical expression of the estimators.

At the end of the paper we present some numerical tests. Due to the inherent difficulty in finding fully published data for this estimation we resort to simulations and real tests on discovering the size of an alphabet. The tests seem to indicate that the new estimator of T is more flexibile than existing ones and thus preferable, in the sense that the performance of all estimators seem to greatly depend on the normalized variance γ^2, and the new estimator is the only one able to perform rather well for all values of $\gamma^2 \in [0,1]$. In our method, the only constraint is that $\lambda \geq 1$, which is $\gamma^2 \leq 1$, which is

imposed in order to ensure convergence of the prior; this, in turn, imposes a mild limitation on the populations to which the method can be applied, since γ^2 can, for some peculiar population, exceed 1; on the other hand, such populations are likely to be quite unusual and, in addition, all other existing estimators seem also to fail on samples taken from them.

In section 2 we review in detail some known estimation methods of interest in deriving our results; in section 3 we derive some relations between known estimators and our first improvements; in section 4 we develop the Bayesian method and define our final estimator; in section 5 we give estimates of the species probabilities from, for both the observed and the unobserved ones; from these, we indicate how to generate confidence intervals for T by means of resampling; finally, in section 6 we present some simulations revealing a rather good performance of our new estimator and also very adequate results of the confidence intervals. All detailed mathematical proof are deferred to the Appendix.

2. Some known estimators of T and related quantities. We start with some notation. Assume that the population from which the sample is drawn has a total of T species (which we sometimes will call states) having proportions $p_1, p_2, \cdots, p_T$.; and that in a sample $x_1, x_2, \cdots, x_n$ of size n there are N observed species. For $i = 1, \cdots, T$, let m_i be the number of observations of the species i in the sample, so that $\sum_{i=1}^{N} m_i = n$. We assume that the m_i's are given one of the possible orders in which $m_1 \geq m_2 \ldots, m_N \geq 1$ and $m_i = 0$ for $i = N+1, \ldots, T$. Also, for $j = 1, \cdots, n$, let n_j be the prevalence of j, which is to say the number of species observed exactly j times, so that $\sum_{j=1}^{n} n_j = N$. Next, let $L_n(i) = m_i/n$ be the empirical frequency of species i, so that $C = \sum_{i:L_n(i)>0} p_i$ is the coverage, i.e, the total probability of the observed species, and $U = 1 - C = \sum_{i:L_n(i)=0} p_i$ is the unobserved probability. We are interested in the estimation of T from the prevalences.

The estimation of U has also been studied intensively (see, for instance, [40] and [38]). In fact, it is possible to turn the estimation of U into a simplified version of our original problem by assuming that there are $N + 1$ species, the N observed ones and the "new" species with probability U; the main issue becomes then the estimation of the probabilities of the various species and especially for the new one. For this and other reasons that we shall see, the estimations of T and U are closely intertwined (even the title of [20] points to this relation).

The first attempt to estimate U can be extracted from Laplace (see [34] and [45]) who suggested an "add-one" estimator: this consists in adding

one to the number of observations of each species plus an additional one for the "unobserved" species. In an extended version, which can be named "add λ", one can add some positive value λ to each species' number of observations (including the unobserved one): an estimate of the probability of each observed species i is then $\widehat{p_i} = \frac{m_i+\lambda}{\lambda+\sum_{i\geq 0}(m_i+\lambda)} = \frac{m_i+\lambda}{n+(N+1)\lambda}$ and the estimate of the unobserved probability becomes $\widehat{U}_{L,\lambda} = \frac{\lambda}{n+(N+1)\lambda}$.

With a seemingly completely different method, Turing and Good (see [24]) proposed another estimator of U. Recall that n_1 is the number of species observed exactly once and n the size of the sample; then the Turing-Good estimator for U is some minor modification of:

$$\widehat{U}_{TG} = \frac{n_1}{n}.$$

A plausible rationale for this estimator is that while for species observed at least twice the empirical frequency is already becoming stable and very likely close to the corresponding probability, species observed only once are likely to be randomly selected representatives of the collection of the yet unobserved species. A more sound mathematical derivation is in Good ([24]), in which also a "'smoothing"' of the n_i's is proposed.

Other methods to estimate U have been developed, and in particular we refer to [38] for a Bayesian method based on the general class of Gibbs-type priors (see also [46] and the other references in [38] for the definition and properties of such priors). This class contains several known families of priors as particular cases and each such family is based on one or more parameters, which need to be further estimated. In [38], for instance, a maximum likelihood estimator is used. Another recent advance appears in Orlitsky et al ([45]), in which a quantity is introduced, called attenuation, that measures the effectiveness of the estimation of U as the sample gets larger; the performance of an estimator is compared to the maximum probability of the observed prevalences and asymptotically very good estimators are determined.

We are going to base our work here on a preliminary estimation of U. It is conceivable that within the wide class of proposed estimators of U some would improve the results that we get; however, we focus on the unsmoothed Turing-Good estimator since it is more direct and simple, while still allowing us to achieve very satisfactory results.

Getting back to the estimation of T, there are several parametric methods based on assuming some structure of the species distribution; for instance,

an estimator devised for the uniform case, in which the probabilities of all species are assumed to be the same is the Horvitz-Thompson

$$\widehat{T}_{HT} = \frac{N}{1 - U},$$

(see [39] and Bishop, Fienberg and Holland (1975) [3]) and then U can be further estimated, for instance by the unsmoothed Turing-Good method, to get

$$(1) \qquad \widehat{T}_{HTTG} = \frac{N}{1 - \widehat{U}_{TG}} = \frac{nN}{n - n_1}$$

see [16] and [5]. Esty [20] improves this estimate by assuming a negative binomial prior with parameter k to get

$$(2) \qquad \widehat{T}_{HTTG} = \frac{N}{1 - \widehat{U}_{TG}} + \frac{n\widehat{U}_{TG}}{(1 - \widehat{U}_{TG})}\frac{1}{k},$$

then providing some ad hoc guess for k (in some cases, $k = 2$).

As to nonparametric methods, Harris [28], Chao [12] and Chao & Lee [13] have proposed some such estimators, of which the most reliable ones seem to be those proposed in [13]. In our notation these amount to

$$(3) \qquad \widehat{T}_{CL}(\widehat{\gamma}) = \frac{N}{1 - \widehat{U}_{TG}} + \frac{n\widehat{U}_{TG}}{(1 - \widehat{U}_{TG})}\widehat{\gamma}^2,$$

with $\widehat{\gamma}^2$ an estimate - for which Chao & Lee make two proposals - of the normalized variation coefficient of the p_i's. In fact, assume that p is a random variable uniformly distributed on the T population probabilities $p_1, \ldots, p_T$;, then its average is

$$\bar{p} = \frac{1}{T}\sum_{k=1}^{T} p_k = \frac{1}{T},$$

and its normalized variation coefficient is

$$(4) \qquad \gamma^2 = \frac{Var(p)}{[\mathbb{E}(p)]^2} = T\sum_{k=1}^{T}(p_k - \bar{p})^2 = T\sum_{k=1}^{T} p_k^2 - 1.$$

Next, Chao and Lee proceed by using an estimate of Good and Toulmin

$$(5) \qquad \sum_{k=1}^{T} p_k^2 \approx \widehat{V}_{GT} = \sum_{j \geq 1} \frac{j(j-1)n_j}{n(n-1)}$$

and using one preliminary estimate for T, (1) for instance, to obtain

$$\widehat{\gamma}^2 = \max\Big(\frac{nN}{n - n_1}\sum\frac{j(j-1)n_j}{n(n-1)} - 1, 0\Big).$$

Note that the work by Chao and Lee can be considered as a further improvement over the results by Esty. However, Chao and Lee make a rather direct use of a preliminary guess for T and we think their method is too sensitive to errors in such preliminary evaluation. In the next section we start discussing some possible improvements.

3. Preliminary results on new estimators. (I) We first consider (3) and (4) as equations in the unknowns T and γ^2 and search for simultaneous solutions $T \geq N$ and $\gamma^2 \geq 0$. Since in some simple examples the unique solution gives $\gamma^2 < 0$, we consider the solutions $T_1(\widehat{\gamma}_1^2)$ and $\widehat{\gamma}_1$ of the problem

(6) $$T \;=\; T(\gamma^2) = \frac{N}{1 - \widehat{U}_{TG}} + \frac{n\widehat{U}_{TG}}{(1 - \widehat{U}_{TG})}\gamma^2$$

(7) $$\widehat{\gamma}^2 \;=\; \arg\inf_{\gamma^2 \geq 0}\left|\gamma^2 - (T\widehat{V}_{GT} - 1)\right|,$$

with $\widehat{V}_{GT}$ as in (5). On letting $u = \widehat{U}_{TG}$ and $v = \widehat{V}_{GT}$ for brevity, the function to minimize becomes

$$(1 - u + nuv)\gamma^2 + 1 - u - Nv;$$

note that $(1 - u + nuv) \geq 0$ since $u \leq 1$, so that the solutions of (6) are

$$\widehat{\gamma}_1^2 = \begin{cases} 0 & \text{if } 0 < u \leq 1 - Nv \\ \frac{Nv-1+u}{1-u+nuv} = \frac{N\widehat{V}_{GT}-1+\widehat{U}_{TG}}{1-\widehat{U}_{TG}+n\widehat{U}_{TG}\widehat{V}_{GT}} & \text{if } 1 - Nv < u \end{cases}$$

and $\widehat{T}_1 = T_1(\widehat{\gamma}_1^2)$.

Some tests described in section 6 show that $\widehat{T}_1$ performs better for non uniform populations than the original Chao-Lee estimate, but has too large a variance.

(II) Next we compare two estimators of U, the unsmoothed Turing-Good and the following modified version of the "add λ": assume the number T of species is known and add λ to each of the frequencies of all the T species,

not just to that of those arbitrarily labelled through $N + 1$. This would give

$$\widehat{p}_k(\lambda) = \frac{m_k + \lambda}{T\lambda + n} \quad \text{per } k = 1 \dots N$$

$$\widehat{p}_k(\lambda) = \frac{\lambda}{T\lambda + n} \quad \text{per } k = N + 1 \dots T$$

$$\widehat{U}_\lambda = \frac{(T - N)\lambda}{T\lambda + n}$$

since there are $T - N$ unobserved species. Now, we can hope to reconcile the extended "add λ" and the unsmoothed Turing-Good estimators by requiring that they assign the same value to $\widehat{U}$. This amounts to solving

(8)
$$\frac{(T - N)\lambda}{T\lambda + n} = \widehat{U}_{TG} = \frac{n_1}{n}.$$

Solving for T we get

(9)
$$\widehat{T}_\lambda = \frac{N + n\widehat{U}_{TG}/\lambda}{1 - \widehat{U}_{TG}} = n\frac{N + n_1/\lambda}{n - n_1}.$$

Quite surprisingly, we have obtained

LEMMA 3.1. *The only value of T for which the extended "'add λ"' and the Turing-Good estimators of U coincide, is the Chao-Lee estimator $T_{CL}(\gamma)$ with $\gamma^2 = 1/\lambda$. From now on we will assume this equality and mostly refer to the parameter λ.*

(III) The relation found in (II) suggests that (6) can be seen as a first moment estimate:

(10)
$$\sum_{k=N+1}^{T} \widehat{p}_k(\lambda) = \widehat{U}_{TG},$$

so that one can hope to derive γ^2 from a second moment relation. The form is suggested by (I), considering the meaning of $\widehat{V}_{GT}$:

(11)
$$\widehat{\lambda}_2 = \arg\inf_{\lambda \geq 0} |\sum_{k=1}^{T} \widehat{p}_k(\lambda)^2 - \widehat{V}_{GT}|.$$

The solutions $\widehat{T}_2(\widehat{\lambda}_2)$ and $\widehat{\lambda}_2$ of (10) and (11), together with $\widehat{\gamma}^2 = \widehat{\lambda}_2^{-1}$, give new estimators; although this seems to improve the estimation in some cases, it does appear to have significant flaws, as shown in the simulations reported in tables 1-3.

4. The Bayesian interpretation. To further improve the above estimate, we need to understand more about the "add λ" estimator. It turns out, as was probably known already to Laplace, that the probability estimation according to the "add λ" method is nothing but the average species probability under the Bayesian posterior on probability distributions on T species

$$\Sigma_T = \{p = (p_1, p_2, \cdots, p_T), p_i \geq 0, \sum_{i=1}^{T} p_i = 1\},$$

given the sample, with a single parameter Dirichlet prior $\rho_{0,T,\lambda}$, i.e. a prior with density $c\prod_{i=1}^{T} p_i^{\lambda-1}$ for some constant c and $\lambda \geq 1$. With likelihood

$$\mu(x) = c\prod_{j=1}^{n} p_{x_j}^{\lambda-1} = c\prod_{i=1}^{T} p_i^{m_i+\lambda-1}$$

the posterior becomes

$$(12) \qquad \rho_{n,T,\lambda}(d\mu) \quad = \quad \frac{\mu(x)\rho_{0,T,\lambda}(d\mu)}{\int_{\Sigma_T} \mu(x)\rho_{0,T,\lambda}(d\mu)}$$

$$= \quad \rho_{n,T,\lambda}(d\mu) = \frac{1}{Z_\Lambda}\mathbf{1}_{\Sigma_T}\prod_{i=1}^{T} p_i^{m_i+\lambda-1}dp_1\ldots dp_T.$$

where $Z = \int_{\Sigma_T} p_1^{m_1+\lambda-1}\cdots p_N^{m_N+\lambda-1}p_{N+1}^{\lambda-1}\cdots p_T^{\lambda-1}dp_1\cdots dp_T$ (note that the constant terms have been cancelled).

By standard integration using the gamma function (see Appendix 1), we find that the average species probability under the posterior is:

$$E_{\rho_{n,T,\lambda}}(y_i) = \begin{cases} \frac{m_i+\lambda}{T\lambda+n} & \text{if } i = 1, \ldots, N \\ \frac{\lambda}{T\lambda+n} & \text{if } i = N+1, \ldots, T \end{cases}$$

as claimed. This remark, together with our reconciliation Lemma in (I) above, indicates that we are taking a new step in the development which brought us from (1) to (2) and then to (3) by assigning now two other meanings for $\lambda^{-1} = \gamma^2$, namely that of the add constant in a generalized Laplace method and that of the constant in a Dirichlet prior.

The Bayesian interpretation of $\widehat{p}_k$ also suggests a modification of the second moment minimization (11). Recalling that now $\lambda \geq 1$ we have:

$$\widehat{\lambda} = \arg\inf_{\lambda \geq 1} |f(\lambda)|$$

with

$$f(\lambda) = \widehat{V} - \sum_{k=1}^{T}(\mathbb{E}_{\rho_{n,T,\lambda}}(p_k^2))$$

$$= \frac{\sum_{j>0} j^2 n_j - n}{n(n-1)} - \frac{2n\lambda + \lambda + n\lambda(\lambda+1)\frac{N\lambda+n_1}{n-n_1}}{[n\frac{N\lambda+n}{n-n_1}+1][n\frac{N\lambda+n}{n-n_1}]}$$

where $\widehat{T}_\lambda$ has been taken as in (9) and the calculation is carried out in Appendix 1. In Appendix 2 we show the function $f(\lambda)$ has two singularities $\beta_2 < \beta_1 = -\frac{n}{N} < 0$ and two zero's, the interesting one being

$$(13) \qquad \lambda_2 = \frac{1 - u - v + uv - uvn}{Nv + u - 1}.$$

The minimization depends on the sign of $f(\lambda)$ for large λ which in turn depends on the sign of $(\lambda_2 - \beta_1)$. Since $f(\lambda)$ is increasing for $\lambda \geq 1$, if the limit for large λ is negative, then the only reasonable value we can assign is ∞, else there is a real solution for the minimization problem above: note that if $\lambda_2 \leq 1$ then we are forced to take $\widehat{\lambda} = 1$. It is thus shown in Appendix 2 that the minimization above yields the estimator

$$\widehat{\lambda} = \begin{cases} 1 & \text{if } \beta_1 < \lambda_2 \text{ and } 1 \geq \lambda_2, \text{ i.e. } \frac{2-v(N+1)}{2-v+vn} \leq u \leq 1-v \\ \lambda_2 & \text{if } \beta_1 < \lambda_2 \text{ and } \lambda_2 \geq 1, \text{ i.e. } 1 - Nv < u \leq \frac{2-v(N+1)}{2-v+vn} \\ \infty & \text{if } \lambda_2 \leq \beta_1, \text{ i.e. } 0 \leq u \leq 1 - Nv. \end{cases}$$

From (9) we get the following estimator of T:

$$\widehat{T}_{\widehat{\lambda}} = \frac{N + n\widehat{U}_{TG}/\widehat{\lambda}}{1 - \widehat{U}_{TG}} = \begin{cases} n\frac{N+n_1/(\lambda_2 \vee 1)}{n-n_1} & \text{if } \beta_1 < \lambda_2 \\ \frac{nN}{n-n_1} & \text{if } \lambda_2 < \beta_1. \end{cases}$$

or, alternatively,

$$\widehat{T}_{\widehat{\lambda}} = \begin{cases} \frac{N+nu}{1-u} & \text{if } \frac{2-v(N+1)}{2-v+vn} \leq u \leq 1-v \\ \frac{N-Nv-nu}{1-u-v+uv-uvn} & \text{if } 1 - Nv \leq u \leq \frac{2-v(N+1)}{2-v+vn} \\ \frac{N}{1-u} & \text{if } 0 \leq u \leq 1 - Nv. \end{cases}$$

Clearly, $\widehat{T}_{\widehat{\gamma}^2}$ is not necessarily an integer while T is such, and we round it to the nearest integer. Notice that when $\widehat{\lambda} = \infty$ we get $\widehat{T}_{\widehat{\lambda}} = \widehat{T}_{HTTG}$.

**5. Estimate of species distribution and confidence intervals for
T.**   Since we now have an estimate for both the parameters T and λ, we
can use the posterior average probability of each species as an estimate of
the species probabilities. For the observed species, i.e. for $i = 1, \dots, N$, this
amounts to

$$(14) \qquad \widehat{p}_i = E_{\rho_{n,\widehat{T}_{\widehat{\lambda}},\widehat{\lambda}}}(y_i) = \frac{m_i + \widehat{\lambda}}{\widehat{T}_{\widehat{\lambda}}\widehat{\lambda} + n} = \frac{(m_i + \widehat{\lambda})(1 - \widehat{U})}{n + N\widehat{\lambda}}.$$

This expression is correct also for $\widehat{\lambda} = \infty$ in which case all species are es-
timated to have probability $(\widehat{T})^{-1}$. Also note that these values are close to
the unbiased estimator m_i/n of the probability of the i-th species and can
be seen as a mixture of the Laplace add-λ and Turing-Good estimators since
they are obtained by adding λ to the frequency m_i of the N observed species
(recall that $n = \sum_{i=1}^{N} m_i$), but only after having assigned the probability
$\widehat{U}$ to the event that we will observe a new species; the estimate of each of
the N species is then reduced by the factor $1 - \widehat{U}$ to compensate for this
and, in fact, $(\widehat{T}_{\widehat{\lambda}} - N)\frac{\widehat{\lambda}(1 - \widehat{U})}{n + N\widehat{\lambda} + \widehat{U}} = \widehat{U}$. This is likely to be a sensible way to
make the attenuation of the Laplace estimator (see [45]) finite. An alterna-
tive description of our estimator is then completed by using the previously
estimated value of λ.

A simple approach for the unobserved species would be to uniformly split
the probability $\widehat{U}$ among the $\widehat{T}_{\widehat{\lambda}} - N$ unobserved species and by the reconcila-
tion method in (8) and (9) this would give $\frac{\widehat{U}}{\widehat{T}_{\widehat{\lambda}} - N} = \frac{\widehat{\lambda}}{\widehat{T}_{\widehat{\lambda}}\widehat{\lambda} + n} = \frac{\widehat{\lambda}(1 - \widehat{U})}{n + N\widehat{\lambda}}$. On the
other hand, notice that, since one can read (10) as $1 - \sum_{k=1}^{N} \widehat{p}_k(\lambda) = 1 - \widehat{U}_{TG}$,
the reconciliation method never used the moments of the p_i's for $i > N$;
therefore, we have some freedom in assigning the estimated values of the
p_i's for $i > N$. These values can then be estimated by taking into account
the meaning of $\lambda^{-1} = \gamma^2$ as normalized species variance, or of some related
quantities; we could then assign probabilities to the unobserved species to
achieve the estimated normalized variance $\widehat{\gamma}^2$ or to achieve some related
equality. For simplicity we will actually focus on $\sum_{k=1}^{N} p_k^2$ and its estimator
$\widehat{V}$. This is a valid approach except when $u < 1 - Nv$, in which case $f(\lambda) < 0$
and $\widehat{V}$ turns out to be too small to be a reasonable estimate of $\sum_{k=1}^{N} p_k^2$; in
that case we replace $\widehat{V}$ with $\sum_{k=1}^{\widehat{T}_{\widehat{\lambda}}} \mathbb{E}_{\rho_{n,\widehat{T}_{\widehat{\lambda}},\widehat{\lambda}}}(p_k^2)$. Clearly

$$\sum_{k=1}^{N}(\mathbb{E}_{\rho_{n,T,\lambda}}(p_k))^2 \le \widehat{V} \vee \sum_{k=1}^{N}\mathbb{E}_{\rho_{n,T,\lambda}}(p_k^2)$$

by Jensen's inequality, and thus we require that the estimates $\widehat{p}_k$ of the probabilities of the unobserved species satisfy:

$$\sum_{k=N+1}^{\widehat{T}_{\widehat{\lambda}}} (\widehat{p}_k)^2 = \left(\widehat{V} \vee \sum_{k=1}^{\widehat{T}_{\widehat{\lambda}}} \mathbb{E}_{\rho_{n,\widehat{T}_{\widehat{\lambda}},\widehat{\lambda}}}(p_k^2) \right) - \sum_{k=1}^{N} (\mathbb{E}_{\rho_{n,\widehat{T}_{\widehat{\lambda}},\widehat{\lambda}}}(p_k))^2 =: \tilde{V}$$

We can use any two parameter distribution, such as for instance $p_i = c\alpha^{i-N}$ for $i = N + 1, \ldots, \widehat{T}_{\widehat{\lambda}}$, and insist that

$$(15) \qquad \sum_{i=N+1}^{\widehat{T}_{\widehat{\lambda}}} p_i = \widehat{U}_{TG}$$

and

$$(16) \qquad \sum_{i=N+1}^{\widehat{T}_{\widehat{\lambda}}} p_i^2 = \tilde{V}.$$

Solving for c and α gives the estimated unobserved probabilities $\widehat{p}_i = p_i(c, \alpha)$, which are used in the simulations of section 6 below to generate confidence intervals by resampling.

It is easily seen that if $T >> N$ then

$$\alpha(1 - \alpha) \approx u/v$$

and

$$c \approx \frac{u(1 - \alpha)}{\alpha}.$$

6. Simulations. In this section we present numerical simulations and tests of the performance of several estimators compared to those we have developed here. Tables 1-4 present the analysis of several populations increasing values of γ^2. Tables 5-6 present some real tests based on discovering the number of letters in an alphabet from a long text. In table 7 we compute confidence intervals using a resampling based on the reconstructed species' probabilities as described in section 5 above.

The estimators compared in tables 1-6 are $\widehat{T}_1$, $\widehat{T}_2$ and $\widehat{T}_{\widehat{\lambda}}$ defined here, then $\widehat{T}_{THTTG}$ from (1), $\widehat{T}_{CL}$ from (3), the Jackknife estimator with optimal parameter $\widehat{T}_{JK}$ from [9] (// indicates numerical errors due to small denominators), and $\widehat{T}_{+1}$ which is our (or the Chao-Lee) estimator with $\gamma^2 = 1$.

In tables 1-4 each population is generated from T i.i.d. random variables, normalized to sum to 1; the resulting γ^2 is determined as normalized interspecies variance; 1000 simple random samples of size n are then generated; finally, mean, SD and mean square error are computed for each estimator.

Tables 5 and 6 test the letter content of some passages in English and Italian in order to detect the number of letters in each alphabet. Each table shows the results of taking 1000 samples of about 9000 letters each from the indicated texts.

The conclusion that can be drawn from these tests is that estimator performances are seen to depend on γ^2, with the $\widehat{T}_{\widehat{\lambda}}$ presenting a consistent low value of the MSE as long as $\gamma^2 \in [0,1]$. Therefore, $\widehat{T}_{\widehat{\lambda}}$ has the flexibility to adapt to the different values of the interspecies variance. In table 1, in fact, $\gamma^2 \approx 0$ and the best estimators turn out to be $\widehat{T}_{THTTG}$ and $\widehat{T}_{CL}$ (in which clearly γ^2 gets appropriately estimated), but all the estimators defined in the present paper perfom equally well. In the less uniform population in table 2, Jackknife and $\widehat{T}_{\widehat{\lambda}}$ show the best performances; and in table 3 where $\gamma^2 \approx 1$, the best estimator turns out to be $\widehat{T}_{+1}$, while $\widehat{T}_{\widehat{\lambda}}$ has only a slightly worse performance. Note that $\widehat{T}_1$ and $\widehat{T}_2$ show a very poor performance in table 2 and 3.

Finally, table 4 shows an extremely skewed population, with γ^2 very large, for which no estimator works properly. The reason for $\widehat{T}_{\widehat{\lambda}}$ is that convergence of the prior imposes $\gamma^{-2} = \lambda \geq 1$.

Even in the alphabet test the performance of $\widehat{T}_{\widehat{\lambda}}$ turns out to be overall best.

Table 7 shows some simulations about confidence intervals for T based on samples of size $n = 400$ computed from $\widehat{T}_{\widehat{\lambda}}$ by estimating the species probabilities p_k as described in section 5 and then resampling 1000 times from the estimated population distribution. This process is repeated 100 times and table 7 indicates, for the populations of tables 1-3 respectively, the percentage of times the confidence intervals hits the true value of $T = 1000$ and the average size of the confidence interval.

The hitting percentage comes out remarkably well, due to the good approximation of the true population distribution by the estimated one.

$T = 1000$	$n = 500$			$n = 1000$			$n = 2000$		
	mean	std	MSE	mean	std	MSE	mean	std	MSE
$\widehat{T}_{TG}$	994	79	79	999	36	36	997	16	16
$\widehat{T}_{CL}$	1010	86	87	1009	42	43	1000	18	18
$\widehat{T}_{JK}$	1068	96	117	1223	84	239	1117	165	203
$\widehat{T}_{+1}$	1759	157	775	1580	73	585	1309	32	311
$\widehat{T}_{1}$	1003	82	82	1005	39	40	1000	18	18
$\widehat{T}_{2}$	1017	83	86	1010	38	40	1027	54	60
$\widehat{T}_{\lambda}$	1087	193	212	1026	60	66	1002	20	20

TABLE 1
Uniform population: p_i's $\sim N(0,1)$, $\gamma^2 \approx 0.009$.

$T = 1000$	$n = 500$			$n = 1000$			$n = 2000$		
	mean	std	MSE	mean	std	MSE	mean	std	MSE
$\widehat{T}_{TG}$	781	59	226	816	30	186	858	15	142
$\widehat{T}_{CL}$	808	72	205	847	40	158	893	20	109
$\widehat{T}_{JK}$	962	88	96	1034	88	94	1054	763	764
$\widehat{T}_{+1}$	1342	116	361	1245	59	252	1118	30	122
$\widehat{T}_{1}$	796	64	213	835	35	168	884	19	117
$\widehat{T}_{2}$	787	59	220	816	30	186	858	15	142
$\widehat{T}_{\lambda}$	915	189	207	891	65	127	912	26	92

TABLE 2
Less uniform population: p_i's $\sim U[0,1]$, $\gamma^2 \approx 0.3317$.

7. APPENDIX 1: The Bayesian approach. By definition of the gamma and beta functions $\Gamma(x) = \int_0^{+\infty} e^{-t} t^{x-1} dt \quad x > 0$ and

$$\beta(x,y) = \int_0^1 t^{a-1}(1-t)^{b-1} dt = \frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)},$$

taking $z = y/(1-x)$ we get

$$\int_0^{1-x} y^a(1-x-y)^b dy = \int_0^1 (1-x)^{a+b+1} z^a(1-z)^b dz = (1-x)^{a+b+1}\frac{\Gamma(a+1)\Gamma(b+1)}{\Gamma(a+b+2)}.$$

Next, let $\rho_{n,T,\lambda}$ be the Bayesian posterior, given a sample with species records $m_1, \ldots, m_N$, from a Dirichlet prior with parameter λ on

$$Q_T = \{p = (p_1 \ldots p_{T-1}) : p_k > 0, \sum_{k=1}^{T-1} p_k \leq 1\}.$$

$T = 1000$	$n = 500$			$n = 1000$			$n = 2000$		
	mean	std	MSE	mean	std	MSE	mean	std	MSE
$\widehat{T}_{TG}$	620	43	382	659	24	341	759	16	241
$\widehat{T}_{CL}$	690	65	316	784	46	220	888	30	115
$\widehat{T}_{JK}$	870	119	176	955	432	435	1027	661	662
$\widehat{T}_{+1}$	1036	85	92	990	47	48	1013	30	30
$\widehat{T}_1$	658	52	345	733	33	268	845	23	156
$\widehat{T}_2$	620	43	382	659	24	341	759	16	241
$\widehat{T}_\lambda$	910	164	187	973	66	71	1001	42	42

TABLE 3
Non-uniform population: p_i's $\sim Exp(1)$, $\gamma^2 \approx 0.9992$.

$T = 1000$	$n = 500$			$n = 1000$			$n = 2000$		
	mean	std	MSE	mean	std	MSE	mean	std	MSE
$\widehat{T}_{TG}$	192	10	808	228	8	772	261	7	738
$\widehat{T}_{CL}$	262	26	737	346	28	654	416	29	583
$\widehat{T}_{JK}$	326	486	830	//	//	//	//	//	//
$\widehat{T}_{+1}$	271	19	729	304	15	696	334	14	666
$\widehat{T}_1$	231	16	768	291	14	708	344	14	656
$\widehat{T}_2$	192	10	808	228	8	772	261	7	738
$\widehat{T}_\lambda$	271	19	729	304	15	696	334	14	666

TABLE 4
Extremely skewed population: p_i's $\sim \Gamma(1,1)$, $\gamma^2 \approx 9.1289$.

Note that $\rho_{n,T,\lambda}$ is invariant under permutation of the p_k's, so it is valid to express any result via a permutation of indices from a proven statement. Therefore, in the following Theorems it is sufficient to prove the results for some index i.

THEOREM 7.1. *For evey $\lambda \geq 1$ and for every $i = 1 \ldots T$,*

$$(17) \qquad \mathbb{E}_{\rho_{n,T,\lambda}}(p_k) = \frac{m_k + \lambda}{T\lambda + n}.$$

PROOF. For $i \in \{1 \ldots T - 1\}$ we have:

$$
\begin{aligned}
\mathbb{E}_{\rho_{n,T,\lambda}}(p_i) &= \int_{Q_T} p_i \rho_{n,T,\lambda}(d\mu) \\
&= \frac{\int_{Q_T} p_1^{m_1+\lambda-1} \ldots p_i^{m_i+\lambda} \ldots (1 - p_1 - \ldots - p_{T-1})^{m_T+\lambda-1} dp_1 \ldots dp_{T-1}}{\int_{Q_T} p_1^{m_1+\lambda-1} \ldots p_i^{m_i+\lambda-1} \ldots (1 - p_1 - \ldots - p_{T-1})^{m_T+\lambda-1} dp_1 \ldots dp_{T-1}}.
\end{aligned}
$$

$T = 26$	$n = 15$			$n = 25$			$n = 50$		
	media	std	MSE	media	std	MSE	media	std	MSE
$\widehat{T}_{TG}$	19.8	10.3	70	18.6	3.6	56	19.5	2.5	43
$\widehat{T}_{CL}$	22.7	13.3	71	20.7	5.6	56	21.6	4.0	37
$\widehat{T}_{JK}$	19.4	6.9	60	23.0	9.8	58	//	//	//
$\widehat{T}_{+1}$	33.3	20.3	109	27.0	7.0	56	25.6	4.8	29
$\widehat{T}_{\lambda}$	26.9	15.8	87	23.2	7.5	60	23.4	5.3	36

TABLE 5
Estimates for the 26 letters English alphabet from samples drawn from [10]; $\gamma^2 \approx 0.7029$
(see [36])

$T = 21$	$n = 15$			$n = 25$			$n = 50$		
	media	std	MSE	media	std	MSE	media	std	MSE
$\widehat{T}_{TG}$	16.0	6.9	62	15.7	3.4	43	16.7	2.0	31
$\widehat{T}_{CL}$	18.4	9.8	71	17.7	5.4	45	18.5	3.3	27
$\widehat{T}_{JK}$	16.9	6.3	51	19.6	8.2	65	//	//	//
$\widehat{T}_{+1}$	26.3	14.3	113	23.1	6.8	48	21.5	4.0	27
$\widehat{T}_{\lambda}$	21.5	12.2	90	19.9	7.3	49	19.8	4.3	29

TABLE 6
Estimates for the 21 letters Italian alphabet from samples drawn from [2]; $\gamma^2 \approx 0.5932$
(see [48])

For $k = 1 \ldots T$, let

$$
\begin{aligned}
s_k &= m_k + \lambda - 1 \\
\widehat{s}_k &= s_k + \delta(k, i)
\end{aligned}
$$

where δ is the Kronecker delta and for $k = 1 \ldots T - 1$ let

$$
I(k) = \int_{Q_k} p_1^{s_1} \ldots p_k^{s_k} (1 - p_1 - \ldots - p_k)^{s_T + \ldots + s_{k+1} + T - k - 1} dp_1 \ldots dp_k
$$

and

$$
G(k) = \frac{\Gamma(s_k + 1)\Gamma(s_T + \ldots + s_{k+1} + T - k)}{\Gamma(s_T + \ldots + s_{k+1} + s_k + T - k + 1)}
$$

and let $\widehat{I}(k)$ and $\widehat{G}(k)$ be as the quantities without hat but with $\widehat{s}_k$ replacing s_k, so that

$$
\mathbb{E}_{\rho_{n,T,\lambda}}(p_i) = \frac{\widehat{I}(T-1)}{I(T-1)}.
$$

Confidence level	Population $->$	1	2	3
90%	fraction of hits	93%	92%	80%
	average interval size	1115	821	707
95%	fraction of hits	95%	98%	89%
	average interval size	1225	889	827
99%	fraction of hits	97%	100%	98%
	average interval size	1520	1064	977

TABLE 7
Summary of confidence interval performances at the given confidence level from $\widehat{T_{\widehat{\lambda}}}$ by resampling.

Now we have

$$
\begin{aligned}
I(T-1) &= \frac{\Gamma(s_{T-1}+1)\Gamma(s_T+1)}{\Gamma(s_T+s_{T-1}+2)}I(T-2) \\
&= G(T-1)\frac{\Gamma(s_{T-2}+1)\Gamma(s_T+s_{T-1}+2)}{\Gamma(s_T+s_{T-1}+s_{T-2}+3)}I(T-3) \\
&= \prod_{k=1}^{T-1}G(k) = \frac{\Gamma(s_T+1)\dots\Gamma(s_1+1)}{\Gamma(s_T+\dots+s_1+T)}
\end{aligned}
$$

and

$$
\widehat{I}(T-1) = \frac{\Gamma(\widehat{s}_T+1)\dots\Gamma(\widehat{s}_1+1)}{\Gamma(\widehat{s}_T+\dots+\widehat{s}_1+T)} = \frac{\Gamma(s_T+1)\dots\Gamma(s_i+2)\dots\Gamma(s_1+1)}{\Gamma(s_T+\dots+s_1+T+1)}
$$

Therefore,

$$
\begin{aligned}
\mathbb{E}_{\rho_{n,T,\lambda}}(p_i) &= \frac{\Gamma(s_i+2)\Gamma(s_T+\dots+s_1+T)}{\Gamma(s_i+1)\Gamma(s_T+\dots+s_1+T+1)} = \frac{s_i+1}{s_T+\dots s_1+T} \\
&= \frac{m_i+\lambda}{m_1+\dots+m_T+T\lambda} = \frac{m_i+\lambda}{T\lambda+n}.
\end{aligned}
$$

It is easily verified that $\sum_{k=1}^{T}\frac{m_k+\lambda}{T\lambda+n} = 1$.

Moreover, adding these values over the $T-N$ unobserved species we get an estimate of U:

$$
\widehat{U}_{+\lambda} = \mathbb{E}_{\rho_{n,T,\lambda}}(U) = \mathbb{E}_{\rho_{n,T,\lambda}}\Big(\sum_{m_i=0}p_i\Big) = \sum_{i=N+1}^{T}\mathbb{E}_{\rho_{n,T,\lambda}}(p_i) = \frac{(T-N)\lambda}{T\lambda+n}
$$

$\square$

LEMMA 7.1. *For every $\lambda \geq 1$ and $i, j = 1 \ldots T$ such that $i \neq j$,*

$$(18) \qquad \mathbb{E}_{\rho_{n,T,\lambda}}(p_i p_j) = \frac{(m_i + \lambda)(m_j + \lambda)}{(T\lambda + n + 1)(T\lambda + n)}$$

PROOF. Following the proof of Theorem 7.1 let, for $k = 1 \ldots T$,

$$\begin{aligned}
s_k &= m_k + \lambda - 1 \\
\widehat{s}_k &= s_k + \delta(i, k) + \delta(j, k), \qquad i \neq j, \quad 1 \leq i, j \leq T - 1
\end{aligned}$$

Thus

$$\mathbb{E}_{\rho_{n,T,\lambda}}(p_i p_j) = \int_{Q_T} p_1 p_j \rho_{n,T,\lambda}(d\mu) = \frac{\widehat{I}(T - 1)}{I(T - 1)}$$

where

$$\begin{aligned}
I(T - 1) &= \frac{\Gamma(s_T + 1) \ldots \Gamma(s_1 + 1)}{\Gamma(s_T + \ldots + s_1 + T)} \\
\widehat{I}(T - 1) &= \frac{\Gamma(s_T + 1) \ldots \Gamma(s_i + 2) \ldots \Gamma(s_j + 2) \ldots \Gamma(s_1 + 1)}{\Gamma(s_T + \ldots + s_1 + T + 2)}
\end{aligned}$$

Therefore

$$\begin{aligned}
\mathbb{E}_{\rho_{n,T,\lambda}}(p_i p_j) &= \frac{\Gamma(s_i + 2)\Gamma(s_j + 2)\Gamma(s_T + \ldots + s_1 + T)}{\Gamma(s_i + 1)\Gamma(s_j + 1)\Gamma(s_T + \ldots + s_1 + T + 2)} \\
&= \frac{(s_i + 1)(s_j + 1)}{(\sum s_k + T)(\sum s_k + T + 1)} \\
&= \frac{(m_i + \lambda)(m_j + \lambda)}{(T\lambda + n)(T\lambda + n + 1)}
\end{aligned}$$

$\square$

LEMMA 7.2. *For every $\lambda \geq 1$ and for every $k = 1 \ldots T$,*

$$(19) \qquad \mathbb{E}_{\rho_{n,T,\lambda}}(p_k^2) = \frac{(m_k + \lambda + 1)(m_k + \lambda)}{(T\lambda + n + 1)(T\lambda + n)}$$

PROOF. As in Theorem 7.1, for $k = 1 \ldots T$ and $i \in \{1 \ldots T - 1\}$ let

$$\begin{aligned}
s_k &= m_k + \lambda - 1 \\
\widehat{s}_k &= s_k + 2\delta(k, i)
\end{aligned}$$

So, $\mathbb{E}_{\rho_{n,T,\lambda}}(p_i^2) = \int_{Q_T} p_i^2 \quad \rho_{n,T,\lambda}(d\mu) = \frac{\widehat{I}(T-1)}{I(T-1)}$ where

$$
\begin{aligned}
I(T-1) &= \frac{\Gamma(s_T+1)\dots\Gamma(s_1+1)}{\Gamma(s_T+\dots+s_1+T)} \\
\widehat{I}(T-1) &= \frac{\Gamma(s_T+1)\dots\Gamma(s_i+3)\dots\Gamma(s_1+1)}{\Gamma(s_T+\dots+s_1+T+2)}
\end{aligned}
$$

Therefore, for $i = 1, \dots, T-1$,

$$
\begin{aligned}
\mathbb{E}_{\rho_{n,T,\lambda}}(p_i^2) &= \frac{\Gamma(s_i+3)\Gamma(s_T+\dots+s_1+T)}{\Gamma(s_i+1)\Gamma(s_T+\dots+s_1+T+2)} = \frac{(s_i+1)(s_i+2)}{(\sum_{k=1}^T s_k + T)(\sum_{k=1}^T s_k + T + 1)} \\
&= \frac{(m_i+\lambda)(m_i+\lambda+1)}{(T\lambda+n)(T\lambda+n+1)}
\end{aligned}
$$

$\square$

LEMMA 7.3.   *If $q = \sum_{j\geq 0} j^2 n_j = \sum_{k=1}^T m_k^2$ we have*

(20)
$$
\sum_{k=1}^T \mathbb{E}_{\rho_{n,T,\lambda}}(p_k^2) = \frac{q + n(2\lambda+1) + T(\lambda^2+\lambda)}{(T\lambda+n+1)(T\lambda+n)}
$$

PROOF. We have

$$
\begin{aligned}
\sum_{k=1}^T \mathbb{E}_{\rho_{n,T,\lambda}}(p_k^2) &= \sum_{k=1}^T \frac{(m_k+\lambda)(m_k+\lambda+1)}{(T\lambda+n)(T\lambda+n+1)} \\
&= \frac{\sum m_k^2 + n(2\lambda+1) + T(\lambda^2+\lambda)}{(T\lambda+n+1)(T\lambda+n)} \\
&= \frac{q + n(2\lambda+1) + T(\lambda^2+\lambda)}{(T\lambda+n+1)(T\lambda+n)}
\end{aligned}
$$

$\square$

8. APPENDIX 2: Some properties of the function defining λ.

Let $u = \widehat{U}$ and $v = \widehat{V}$. We consider now u and v as free variables satisfying some requirements satisfied by the values that, in fact, $\widehat{U}$ and $\widehat{V}$ take on in our estimation, namely $\widehat{U} = \frac{n_1}{n}$ and $\widehat{V} = \widehat{V}_{GT} = \sum_{j\geq 1} \frac{j(j-1)n_j}{n(n-1)}$.

Also let $q = vn(n-1) + n = \sum_{j>0} j^2 n_j$. Then

LEMMA 8.1.   *For every sample, $\widehat{U} + \widehat{V} \leq 1$*

PROOF. Since $q = \sum_{j>0} j^2 n_j$ and $n = \sum_{j>0} j n_j$ we have that

$$\widehat{U} + \widehat{V} \;\; = \;\; \frac{n_1}{n} + \frac{q - n}{n(n-1)} \leq 1$$

is implied by

$$-nn_1 + n_1 - q + n^2 \;\; = \;\; n_1 + (\sum_{j=1}^{N} j n_j)^2 - \sum_{j=1}^{N} n_j (j n_1 + j^2)$$

$$= \;\; \sum_{j=1}^{N} j^2 (n_j^2 - n_j) + n_1 + \sum_{j,r=1,\ldots,N, j \neq r} j n_j r n_r - n_1 \sum_{j=1}^{N} j n_j$$

$$(21) \qquad\qquad \geq \;\; (n_1^2 - n_1) + n_1 + n_1 \sum_{j=2}^{N} j n_j - n_1^2 \geq 0$$

$\square$

LEMMA 8.2. *For every sample,* $N + \widehat{U} - 1 - n\widehat{U} \geq 0$

PROOF. By definition of $\widehat{U}$ we have

$$N + \widehat{U} - 1 - n\widehat{U} = N - n_1 + \frac{n_1}{n} - 1,$$

then either $n = n_1 = N$ and the right hand side becomes 0, or $N - n_1 \geq 1$ and the relation holds. $\square$

LEMMA 8.3. *For every sample,* $(\widehat{V}nN - \widehat{V}N + N - n)n = qN - n^2 \geq 0$

PROOF. Expressing q, N and n as function of the n_j's we get

$$q \;\; = \;\; \sum_{j>0} j^2 n_j$$

$$N \;\; = \;\; \sum_{j>0} n_j$$

$$n \;\; = \;\; \sum_{j>0} j n_j.$$

Then

$$
\begin{aligned}
qN - n^2 &= \Big(\sum_{j>0} j^2 n_j\Big)\Big(\sum_{k>0} n_k\Big) - \Big(\sum_{j>0} j n_j\Big)^2 \\
&= \Big(\sum_{j>0} j^2 n_j^2 + \sum_{j>0}\sum_{k\neq j} j^2 n_j n_k\Big) - \Big(\sum_{j>0} j^2 n_j^2 + \sum_{j>0}\sum_{k\neq j} j n_j k n_k\Big) \\
&= \sum_{j>0}\sum_{k\neq j}(j^2 - jk) n_j n_k \\
&= \sum_{j>0}\sum_{j<k}(j^2 - jk + k^2 - kj) n_j n_k \\
&= \sum_{j>0}\sum_{j<k}(j - k)^2 n_j n_k \geq 0
\end{aligned}
$$

$\square$

Therefore, in the sequel we assume that u and v satisfy the following relations:

(22) $\qquad\qquad 0 \;\leq\; u \leq 1$

(23) $\qquad\qquad 0 \;\leq\; v \leq 1$

(24) $\qquad\qquad 1 \;\geq\; u + v$

(25) $\qquad\qquad 0 \;\leq\; N + u - 1 - nu$

(26) $\qquad\qquad 0 \;\leq\; vnN - vN + N - n$

THEOREM 8.1. *Let*

$$
f(\lambda) = v - \frac{vn(n-1) + n + n(2\lambda + 1) + \lambda(\lambda + 1)\frac{N + nu/\lambda}{1-u}}{[n + \lambda\frac{N+nu/\lambda}{1-u} + 1][n + \lambda\frac{N+nu/\lambda}{1-u}]}
$$

we have

$$
\widehat{\lambda} = \arg\inf_{\lambda \geq 1}\big|f(\lambda)\big| =
\begin{cases}
1 & \text{if } \beta_1 < \lambda_2 \text{ and } 1 \geq \lambda_2,\ \text{i.e. } \frac{2-v(N+1)}{2-v+vn} \leq u \leq 1-v \\
\lambda_2 & \text{if } \beta_1 < \lambda_2 \text{ and } \lambda_2 \geq 1,\ \text{i.e. } 1 - Nv < u \leq \frac{2-v(N+1)}{2-v+vn} \\
\infty & \text{if } \lambda_2 \leq \beta_1,\ \text{i.e. } 0 \leq u \leq 1 - Nv.
\end{cases}
$$

where $\beta_1 = -\frac{n}{N}$ is the largest singularity of $f(\lambda)$ and

$$
\lambda_2 = \frac{1 - u - v + uv - uvn}{Nv + u - 1}.
$$

PROOF. The equation $f(\lambda) = 0$ has solutions:

$$(27) \qquad \lambda_1 \;\; = \;\; \frac{-2n + nu}{N} \leq 0$$

$$(28) \qquad \lambda_2 \;\; = \;\; \frac{1 - u - v + uv - uvn}{Nv + u - 1}$$

The root λ_1 is always non positive and thus it is not interesting and if

$$(29) \qquad \lambda_2 = \frac{1 - u - v + uv - uvn}{Nv + u - 1} \geq 1,$$

then λ_2 achieves the required minimum.

To evaluate the other cases note that the function $f(\lambda)$ has two poles

$$(30) \qquad \beta_1 \;\; = \;\; -\frac{n}{N}$$

$$(31) \qquad \beta_2 \;\; = \;\; -\frac{n}{N} - \frac{1 - u}{N}$$

and $\lambda_1 < \beta_2 < \beta_1$. Moreover,

$$\lim_{\lambda \to \beta_1^+} f(\lambda) = \infty \cdot \operatorname{sgn}(\frac{(u-1)(vnN - vN + N - n)}{N^2}) = -\infty$$

by (22) and (26), and

$$(32) \qquad \lim_{\lambda \to +\infty} f(\lambda) = \frac{Nv + u - 1}{N}.$$

We now verify that

LEMMA 8.4. $f(\lambda)$ *is increasing for* $\lambda > \beta_1$.

PROOF. Let $f'(\lambda) = \frac{(1-u)g(\lambda)}{(n+N\lambda)(1-u+n+N\lambda)}$. Then

$$(33) \qquad \lim_{\lambda \to \beta_1^+} g(\lambda) = n(1-u)^2(vnN - vN + N - n) > 0$$

by (26). Note that g satisfies

$$
\begin{aligned}
g'(\lambda) \;\; = \;\; & 2N^2(N + u - 1 - nu)\lambda \\
& + 2nN(-1 - n + 2N + u - Nu - Nv + nNv + Nuv - nNuv)
\end{aligned}
$$

with the leading coefficient nonnegative by (25). Therefore, if $\lambda > \beta_1 = -\frac{n}{N}$

$$g'(\lambda) > 2nN(1-u)(vnN - vN + N - n) \geq 0$$

again by (26). Thus $g' > 0$ for all $\lambda > \beta_1$ and, by (33), $g > 0$ for all $\lambda > \beta_1$ and since the other factors in f' are also positive, we have that $f' > 0$ for all $\lambda > \beta_1$ as required. □

Now there are three possibilities.

1. If $u \leq 1 - Nv$ then from (32) and the above Lemma, it follows that $f < 0$ for all $\lambda > \beta_1$ and increasing, thus

$$\widehat{\lambda} = \arg\inf_{\lambda \geq 1} |f(\lambda)| = \arg\max \lambda \geq 1 f = +\infty.$$

2. If $1 - Nv < u <$ then from (29) $\lambda_2 \geq 1$ is equivalent to $u \leq \frac{2-v(N+1)}{2-v+vn}$, in which case $\widehat{\lambda} = \lambda_2$.
3. If $\frac{2-v(N+1)}{2-v+vn} < u$ then $\lambda_2 < 1$ and by the Lemma above

$$\widehat{\lambda} = \arg\inf_{\lambda \geq 1} |f(\lambda)| = \arg\min \lambda \geq 1 f = 1$$

The conditions on u and v are translated into those for λ_2 and β_1 by direct calculation.

□

References.

[1] Almudevar, A., Bhattacharya, R.N. and Sastri, C.C.A. (2000): *Estimating the Probability Mass of Unobserved Support in Random Sampling*, J. Stat. Planning and Inference, **91** , 91-105.

[2] M. Adamo, *La matematica nell'antica Cina*, Osiris, Vol. 15. (1968), pp. 175-195.

[3] Bishop, Y. M. M., Fienberg, S. E., Holland P. W. (1975): *Discrete multivariate analysis: theory and practice*, Cambridge, MIT Press.

[4] Boender, C. G. E. , Rinnoy Kan, A. H. G (1987): *A multinomial Bayesan Approach to the Estimation of Population and Vocabulary Size*, Biometrika **74** No. 4, 849-856.

[5] Böhning, D., Schön, D. (2005): *Nonparametric maximum likelihood estimation of population size based on the counting distribution*, Journal of the Royal Stat. Soc. (C) Appl. Statist. **54**, Part 4, 721-737.

[6] Böhning, D., Suppawattanabe, B., Kusolvisitkul, W., Vivatwongkasem, C (2004): *Estimating the number of drug users in Bangkok 2001: A capture-recapture approach using repeated entries in the list*, Europ. J. of Epidemiology **19**, 1075-1083.

[7] Brose, U., Martinez, M.D., Williams, R. J. (2003): *Estimating species richness: sensitivity to sample coverage and insensitivity to spatial patterns*, Ecology **84** No. 9, 2364-2377.

[8] Bunge, J., Fitzpatrick, M. (1993): *Estimating the number of species: a Review*, J. Amer. Stats. Assn. **88** No. 421, 364-373.

[9] Burnahm, K.P., Overton, W. S. (1979): *Robust estimation of population size when capture probabilities vary among animals*, Ecology **60** No. 5, 927-936.

[10] D. Burton, *The History of Mathematics: An Introduction*, McGraw-Hill, 2003.

[11] Carothers (1993): *Estimating the number of species: a Review*, J. Amer. Stats. Assn. **88** No. 421, 364-373.

[12] Chao, A. (1984): *Nonparametric estimation of the number of classes in a population*, Sc. J. of Stat. **11**, 265-270.

[13] Chao, A., Lee, S-M. (1992): *Estimating the number of classes via sample coverage*, J. Amer.Stat.Assn., **87** No.417, 210-217.

[14] Chuch, K. W. , Gale, W. A. (2006): *Enhanced Good-Turing and Cat-Cal: two new methods for estimating probabilities of english bigrams*, Preprint

[15] Colwell: Estimates. Software Freeware. See http://viceroy.eeb.uconn.edu/estimates

[16] Darroch, J.N., Ratcliff (1980): *A Note on Capture-Recapture Estimation*, Biometrics, **36**, 149-153.

[17] Edwards, W.R, Eberhardt, L.L. (1967): *Estimating cottontail abundance from live trapping data*, J. of Wildlife Manag. **33**, 28-39.

[18] Efron, B. (1981): *Nonparametric standard errors and confidence intervals*, Canadian J. Statist. **9**, 139-172.

[19] Efron, B.,Thisted, R. (1976): *Estimating the number of unseen species: how many words did Shakespeare know?*, Biometrika **63**, 435-467.

[20] Esty, W.W. (1985): Estimation of the Number of Classes in a Population and the Coverage of a Sample, *Mathematical Scientist*, **10**, 41-50.

[21] Esty, W.W. (1986): The size of a coverage, *Numismatic Chronicle*, **146**, 185-215.

[22] Fisher, R.A., Steven Corbet, A., Williams, C.B. (1943): *The relation between the number of species and the number of individuals in a random sample of an animal population*, J. An. Ecol., **12** No. 1, 42-58.

[23] Gandolfi, A., Sastri, C.C.A. (2004): *Nonparametric Estimations about Species not observed in a Random Sample*, Milan J. Math **72**, 81-105.

[24] Good, I. J. (1953): *The population frequencies of species and the estimation of population parameters*, Biometrika **40**, 237-266.

[25] Good, I. J. (1965): *The estimation of probabilities: an essay on modern bayesian method*, Research Monograph No. 30 MIT Press.

[26] Good, I. J. (1967): *A Bayesian significance test for multinomial distributions*, J. Roy. Statist. Soc. Ser. B **29**, 399-431.

[27] Good, I. J. and Toulmin, G. (1956): *The number of new species and the increase in population coverage when a sample is increased*, Biometrika **43**, 45-63.

[28] Harris, B. (1968): *Statistical inference in the classical occupancy problem: unbiased estimation of the number of classes*, J. Amer.Stat.Assn. **63**, 837-847.

[29] Huang, S-P and Weir, B.S. *Estimating the Total Number of Alleles Using a Sample Coverage Method* Genetics 2001 159: 1365-1373

[30] Huand, J. (2006): *Maximum likelihood estimation of Dirichlet distribution parameters*, Manuscript.

[31] Jedynak, B., Khudanpur, S., Yazgan, A. (2005) *Estimating Probabilities from Small Samples* , 2005 Proceedings of the American Statistical Association, Statistical computing section [CD-ROM], Alexandria, VA : American Statistical Association.

[32] Jeffreys, H. (1961): *Theory of probability*, Clarendom Press, Oxford, Third Edition.

[33] Johnson, W. E. (1932): *Probability: the deductive and inductive problems*, Mind **49**,409-423.

[34] Laplace (1995):*Philosophical essays in Probabilities*, Springer Verlag, New York.

[35] Lehmann, E. L. (1983): *Theory of point estimation*, Wiley ed., New York.

[36] R.E. Lewand, *Relative Frequencies of Letters in General English Plain text*, Cryptographical Mathematics.

[37] Lewontin, P., Prout, T. (1956): *Estimation of the different classes in a population*, Biometrics **12**, 211-223.

[38] Lijoi, A, Mena, H. R., Prünster, I. (2007) *Bayesian nonparametric estimation of the probability of discovering new species*. Preprint.

[39] Lindsay, B. G., Roeder, K. (1987): *A unified treatment of integer parameter models*, J. Am. Statist. Ass. **82**, 758-764.

[40] Mao, C.X. (2004): *Predicting the conditional probability of discovering a new class*, Journal of the American Statistical Association, **99**, 1108-1118.

[41] Marchand, J.P. and Schroeck, F.E. (1982): *On the Estimation of the Number of Equally Likely Classes in a Population*, Communications in Statistics, Part A–Theory and Methods, **11**, 1139-1146.

[42] McAllester, D. and Schapire, R.E. (2000): *On the Convergence Rate of Good-Turing Estimators*, Conference On Computing Learning Theory (COLT), 1-6.

[43] McNeil, D. (1973): *Estimating an author's vocabulary*, J. Am. Stat. Ass., **68** No. 341, 92-96.

[44] Norris III, J. L., Pollock, K.H., *Non-parametric MLE for Poisson species abundance models allowing for heterogeneity between species*, Environ. Ecol. Statist., **5** 98), 391-402.

[45] Orlitsky, A., Santhanam, N. P, Zhang, J. (2003): *Always Good Turing: Asimptotically Optimal Probability Estimation*, Science, **302** No. 5644, 427-431.

[46] Pitman, J. (2005): *Combinatorial stochastic processes*, Lecture Notes for the St. Flour Summer School .

[47] Shen, T-J., Chao, A., Lin, C-F. (2003): *Predicting the number of new species in further taxonomic sampling*, Ecology, **84** No. 3 , 798-804.

[48] Simon Singh, *Codici e Segreti*, 1999.

[49] Zabell, S. L.(1982): *W. E Johnson's "Sufficientness" Postulate*, The Annals of Statistics, **10 No. 4** , 1090-1099.

Address of L.Cecconi and A.Gandolfi
Dipartimento di Matematica U. Dini,
Università di Firenze,
Viale Morgagni 67/A, 50134 Firenze, Italy
E-mail: cecconi@math.unifi.it
         gandolfi@math.unifi.it

Address of C.C.A. Sastri
Department of Mathematics and Statistics
Missouri University of Science and Technology
(Formerly University of Missouri - Rolla)
Rolla, MO 65409
USA
E-mail: sastric@mst.edu
URL: